

18 MAR 2004

WIPO

PCT

10/522997
PCT/JP 2004/000971

日 本 国 特 許 庁
JAPAN PATENT OFFICE

30.1.2004

02 FEB 2005

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日
Date of Application: 2003年 1月31日

出 願 番 号
Application Number: 特願2003-023342
[ST. 10/C]: [JP2003-023342]

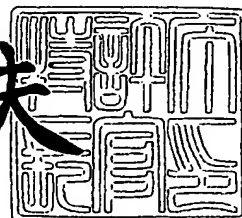
出 願 人
Applicant(s): 松下電器産業株式会社

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN
COMPLIANCE WITH
RULE 17.1(a) OR (b)

2004年 3月 4日

特許庁長官
Commissioner,
Japan Patent Office

今井康夫



出証番号 出証特2004-3016458

【書類名】 特許願

【整理番号】 2033840243

【提出日】 平成15年 1月31日

【あて先】 特許庁長官 殿

【国際特許分類】 G06F 17/00
G06N 5/00
G06N 7/00

【発明者】

【住所又は居所】 大阪府門真市大字門真 1 0 0 6 番地 松下電器産業株式会社内

【氏名】 森川 幸治

【発明者】

【住所又は居所】 北海道札幌市北区北 6 条西 9 丁目 中央公務員宿舎 4 - 1 0 2

【氏名】 大森 隆司

【発明者】

【住所又は居所】 北海道札幌市北区北 1 8 条西 4 丁目 シティハイム N 1 8 - 7 0 8

【氏名】 大東 優

【発明者】

【住所又は居所】 大阪府門真市大字門真 1 0 0 6 番地 松下電器産業株式会社内

【氏名】 岡 夏樹

【特許出願人】

【識別番号】 000005821

【氏名又は名称】 松下電器産業株式会社

【代理人】

【識別番号】 100077931

【弁理士】

【氏名又は名称】 前田 弘

【選任した代理人】

【識別番号】 100094134

【弁理士】

【氏名又は名称】 小山 廣毅

【選任した代理人】

【識別番号】 100110939

【弁理士】

【氏名又は名称】 竹内 宏

【選任した代理人】

【識別番号】 100110940

【弁理士】

【氏名又は名称】 嶋田 高久

【選任した代理人】

【識別番号】 100113262

【弁理士】

【氏名又は名称】 竹内 祐二

【選任した代理人】

【識別番号】 100115059

【弁理士】

【氏名又は名称】 今江 克実

【選任した代理人】

【識別番号】 100115691

【弁理士】

【氏名又は名称】 藤田 篤史

【選任した代理人】

【識別番号】 100117581

【弁理士】

【氏名又は名称】 二宮 克也

【選任した代理人】

【識別番号】 100117710

【弁理士】

【氏名又は名称】 原田 智雄

【選任した代理人】

【識別番号】 100121500

【弁理士】

【氏名又は名称】 後藤 高志

【選任した代理人】

【識別番号】 100121728

【弁理士】

【氏名又は名称】 井関 勝守

【手数料の表示】

【予納台帳番号】 014409

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 0217869

【ブルーフの要否】 要

【書類名】 明細書

【発明の名称】 予測型行動決定装置および行動決定方法

【特許請求の範囲】

【請求項 1】 所定の環境について状態を観察し、状態値を取得する状態観察部と、

前記状態観察部によって取得された状態値に基づいて、前記環境の将来の状態変化を予測する環境予測部と、

前記環境予測部による予測結果を基にして、行動決定のために最も適した将来の状態を目標状態として決定する目標状態決定部と、

前記目標状態決定部によって決定された目標状態を基にして、自己の行動を決定する第 1 の行動決定部とを備えた
ことを特徴とする予測型行動決定装置。

【請求項 2】 請求項 1 において、

前記環境予測部は、前記環境の将来の、自己の行動に影響されない状態変化を予測するものである
ことを特徴とする予測型行動決定装置。

【請求項 3】 請求項 1 において、

前記環境の各状態に係る状態価値を記憶する状態価値記憶部を備え、
前記環境予測部は、複数ステップにわたって、将来の状態を予測するものであり、
前記目標状態決定部は、前記環境予測部によって予測された各将来状態について、それぞれ、前記状態価値記憶部を参照して状態価値を求め、目標状態を決定するものである
ことを特徴とする予測型行動決定装置。

【請求項 4】 請求項 3 において、

前記目標状態決定部は、状態価値が極大となる将来状態を、目標状態として決定する
ことを特徴とする予測型行動決定装置。

【請求項 5】 請求項 3 において、

前記目標状態決定部は、前記状態価値記憶部から得た状態価値を、現在からのステップ数に応じて割り引いて、用いることを特徴とする予測型行動決定装置。

【請求項 6】 請求項 3 において、

前記状態価値記憶部は、自己を含む状態に係る状態価値を記憶するものであり、

当該予測型行動決定装置は、

前記環境予測部によって予測された、自己を含まない将来状態について、前記状態価値記憶部に記憶された状態価値を基にしてその状態価値を求め、前記目標状態決定部に与える価値変換部を備えたものであることを特徴とする予測型行動決定装置。

【請求項 7】 請求項 1 において、

所定の行動基準に基づいて、自己の行動を決定する第 2 の行動決定部と、

前記第 1 および第 2 の行動決定部によって決定された行動を第 1 および第 2 の行動候補として受け、これら第 1 および第 2 の行動候補のうちのいずれか一方を、実際の行動として選択する行動選択部とを備えたことを特徴とする予測型行動決定装置。

【請求項 8】 請求項 7 において、

前記目標状態決定部は、目標状態を決定できたか否かを示す選択信号を前記行動選択部に与えるものであり、

前記行動選択部は、前記選択信号が、目標状態を決定できたことを示すときは、前記第 1 の行動候補を選択する一方、目標状態を決定できなかったことを示すときは、前記第 2 の行動候補を選択するものであることを特徴とする予測型行動決定装置。

【請求項 9】 請求項 1 において、

前記第 1 の行動決定部は、

前記状態値を受け、この状態値が表す現在状態から、その前ステップにおける状態と行動を検出する行動付状態変化検出部と、

前記行動付状態変化検出部によって検出された、現在状態並びにその前ステッ

プにおける状態および行動の組合せを、状態変化として記憶する行動付状態変化記憶部と、

前記行動付状態変化記憶部から、現在状態から目標状態までの期間の状態変化の履歴を検索し、この検索結果を基にして、行動を決定する行動計画部とを備えたものである

ことを特徴とする予測型行動決定装置。

【請求項 10】 請求項 9 において、

前記行動計画部は、前記状態変化記憶部の検索の際に、目標状態から現在の状態に向かって後ろ向き探索を行うものである

ことを特徴とする予測型行動決定装置。

【請求項 11】 請求項 1 において、

前記環境予測部は、

前記状態値を受け、この状態値が表す現在状態から、その前ステップにおける状態を検出する状態変化検出部と、

前記状態変化検出部によって検出された、現在状態およびその前ステップにおける状態の組合せを、状態変化として記憶する状態変化記憶部と、

前記状態変化記憶部から、現在状態の後の状態を予測する状態予測部とを備えたものである

ことを特徴とする予測型行動決定装置。

【請求項 12】 予測型行動決定装置において、自己の行動を決定する方法であって、

所定の環境について状態を観察して、状態値を取得し、

取得した状態値に基づいて、前記環境の将来の状態変化を予測し、

予測結果を基にして、行動決定のために最も適した将来の状態を目標状態として決定し、

決定した目標状態を基にして、自己の行動を決定することを特徴とする行動決定方法。

【請求項 13】 請求項 12 において、

予測する状態変化は、前記環境の将来の、自己の行動に影響されない状態変化

である

ことを特徴とする行動決定方法。

【請求項 14】 請求項 12 において、

目標状態の決定を、予測した各将来状態に係る状態価値を参照して、行うことを特徴とする行動決定方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、外部から入力を受け、現在の状態から将来どのような状態に遷移するかを予測しながら、外部への出力を決定する行動決定に関する技術に属する。

【0002】

【従来の技術】

近年、産業上使用されるシステムは、年々複雑化しており、入力と出力との関係を予めプログラム等によって記述しておくことが困難になりつつある。このため、入力信号を処理して正しい出力を求める方法が必要になっており、このような入力から出力を決定する装置のことを、本願明細書中で「行動決定装置」と呼ぶ。また、特に、入力信号から将来の状態遷移を予測した上で、出力を求めるものを「予測型行動決定装置」と呼ぶ。

【0003】

行動決定のための従来技術は、1) 現在の状態のみによって行動決定を行うもの、2) 過去の状態の遷移によって行動決定を行うもの、3) 将来の状態を予測して行動決定を行うもの、に分類される。

【0004】

現在の状態のみによって行動決定を行う従来技術として、IF-THENルールを用いるもの、ニューラル・ネットワークを用いるもの、テーブル参照方式などがある。これらは、現在の状態に対する行動が予め記述されており、入力から現在の状態を判別し、所定の記述を参照して行動を決定する。

【0005】

ところが、現在の状態だけで正しい行動が決定できるとは必ずしも限らない。

例えば対話型ロボットシステムにおいて、「いいですか」と聞かれたとき、それだけでは意味不明であり、それまでの状態遷移を参照することによって始めて言葉の意味を理解することができる。すなわち、行動決定のために、過去の状態が必要となる場合もある。

【0006】

また、現在や過去の状態だけでなく、将来の状態も考慮すべき場合もある。例えば、移動ロボットが障害物を回避するような場合には、まだ衝突していない段階では問題は発生していなくて、将来このままの移動方向や速度では衝突するという将来の状態遷移を考慮することによって、はじめて衝突前に回避の行動をとることができる。

【0007】

将来の状態を考慮した行動決定に関する従来技術として、特許文献1, 2に開示された技術がある。特許文献1では、環境から視覚センサや関節角度センサによって得た入力信号から、現在の画像データや関節角度データを状態として受ける。システムは、対象物に対する行動とその行動結果としての画像の変化をリカレント・ニューラル・ネットワークによって記憶し、同じような状況を受けた場合に、記憶している行動を再現する。この技術は、例えばロボットの自律的行動決定に応用されている。また特許文献2では、強化学習における行動決定技術が示されており、ある状態における価値と1ステップ前の状態における価値から、誤差を予測してその情報を行動決定に用いている。

【0008】

【特許文献1】

特開 2002-59384号公報

【0009】

【特許文献2】

特開 2002-189502号公報

【0010】

【発明が解決しようとする課題】

上述のように、特許文献1では、現在状態から自己の行動によってどのような

状態に変化するかをリカレント・ニューラル・ネットワークを用いて予測している。そして、その予測結果に応じて、状態と対で記憶された行動が決定される。

【0011】

しかしながら、特許文献1では、自己の行動に起因する過去の状態変化をリカレント・ニューラル・ネットワークを用いて学習しているに過ぎず、自己の行動と関係のない環境の変化については、予測も考慮も何らなされていない。また、ある時点での行動決定は、現在の状態とその1ステップ後の状態予測に基づいて行われているが、この1ステップ後の状態が行動決定のために重要であるとは必ずしもいえず、したがって、この将来の状態予測が行動決定にとって適切なものとはいえない。

【0012】

また、特許文献2に関しても、現在の状態と1ステップ後の状態の予測値のみから決定された行動が、必ずしも望ましいものとは限らない。例えば、移動ロボットが走ってくる車をよけたいときに、移動ロボットの速度が車に対してかなり遅い場合は、何ステップも前から回避行動を起こさないと車とぶつかってしまうように、将来を見越して行動を決定すべき場合には、1ステップ後だけでなくさらに将来の状態も考慮して行動を決定する必要がある。また、例えば上述の移動ロボットが走ってくる車をよける場合のように、現在や1ステップ後の価値を見ても変化はみられないが、何ステップも先に決定的な自体が発生するときは、現在や1ステップ後の価値に基づく行動は、無駄な行動につながる可能性もある。

【0013】

前記の問題に鑑み、本発明は、予測型行動決定装置において、行動決定のために将来の状態予測をより適切に行い、行動決定の精度や能力を向上させることを課題とする。

【0014】

【課題を解決するための手段】

本発明は、次の点に着目している。すなわち、環境の状態の中には、自己（予測型行動決定装置）の行動（＝出力）に関連するものと、自己の行動に関係なく

変化するものがあり、前者については、長い将来にわたっての状態予測は困難であるが、その一方で、後者については、1ステップ先だけではなく長い将来にわたっての予測も比較的容易に行うことができる。

【0015】

すなわち、本発明では、環境を観察して、環境の将来の状態変化を予測し、この予測結果を基にして、行動決定のために最も適した将来の状態を目標状態として決定する。そして、この決定した目標状態を基にして、自己の行動を決定する。これにより、環境の将来の状態予測により得られた、行動決定のために最も適した目標状態を基にして、行動が決定されるので、従来よりも行動決定の精度が向上する。

【0016】

【発明の実施の形態】

本発明の第1態様によれば、所定の環境について状態を観察し、状態値を取得する状態観察部と、前記状態観察部によって取得された状態値に基づいて、前記環境の将来の状態変化を予測する環境予測部と、前記環境予測部による予測結果を基にして、行動決定のために最も適した将来の状態を目標状態として決定する目標状態決定部と、前記目標状態決定部によって決定された目標状態を基にして、自己の行動を決定する第1の行動決定部とを備えた予測型行動決定装置を提供する。

【0017】

本発明の第2態様によれば、前記環境予測部は、前記環境の将来の自己の行動に影響されない状態変化を予測する第1態様の予測型行動決定装置を提供する。

【0018】

本発明の第3態様によれば、前記環境の各状態に係る状態価値を記憶する状態価値記憶部を備え、前記環境予測部は、複数ステップにわたって将来の状態を予測するものであり、前記目標状態決定部は、前記環境予測部によって予測された各将来状態について、それぞれ、前記状態価値記憶部を参照して状態価値を求め、目標状態を決定する第1態様の予測型行動決定装置を提供する。

【0019】

本発明の第4態様によれば、前記目標状態決定部は、状態価値が極大となる将来状態を目標状態として決定する第3態様の予測型行動決定装置を提供する。

【0020】

本発明の第5態様によれば、前記目標状態決定部は、前記状態価値記憶部から得た状態価値を現在からのステップ数に応じて割り引いて用いる第3態様の予測型行動決定装置を提供する。

【0021】

本発明の第6態様によれば、前記状態価値記憶部は、自己を含む状態に係る状態価値を記憶するものであり、当該予測型行動決定装置は、前記環境予測部によって予測された、自己を含まない将来状態について、前記状態価値記憶部に記憶された状態価値を基にしてその状態価値を求め、前記目標状態決定部に与える価値変換部を備えた第3態様の予測型行動決定装置を提供する。

【0022】

本発明の第7態様によれば、所定の行動基準に基づいて自己の行動を決定する第2の行動決定部と、前記第1および第2の行動決定部によって決定された行動を第1および第2の行動候補として受け、これら第1および第2の行動候補のうちのいずれか一方を実際の行動として選択する行動選択部とを備えた第1態様の予測型行動決定装置を提供する。

【0023】

本発明の第8態様によれば、前記目標状態決定部は、目標状態を決定できたか否かを示す選択信号を前記行動選択部に与えるものであり、前記行動選択部は、前記選択信号が、目標状態を決定できたことを示すときは、前記第1の行動候補を選択する一方、目標状態を決定できなかったことを示すときは、前記第2の行動候補を選択する第7態様の予測型行動決定装置を提供する。

【0024】

本発明の第9態様によれば、前記第1の行動決定部は、前記状態値を受け、この状態値が表す現在状態からその前ステップにおける状態と行動を検出する行動付状態変化検出部と、前記行動付状態変化検出部によって検出された、現在状態並びにその前ステップにおける状態および行動の組合せを、状態変化として記憶

する行動付状態変化記憶部と、前記行動付状態変化記憶部から、現在状態から目標状態までの期間の状態変化の履歴を検索し、この検索結果を基にして、行動を決定する行動計画部とを備えた第1態様の予測型行動決定装置を提供する。

【0025】

本発明の第10態様によれば、前記行動計画部は、前記状態変化記憶部の検索の際に、目標状態から現在の状態に向かって後ろ向き探索を行う第9態様の予測型行動決定装置を提供する。

【0026】

本発明の第11態様によれば、前記環境予測部は、前記状態値を受け、この状態値が表す現在状態からその前ステップにおける状態を検出する状態変化検出部と、前記状態変化検出部によって検出された現在状態およびその前ステップにおける状態の組合せを状態変化として記憶する状態変化記憶部と、前記状態変化記憶部から現在状態の後の状態を予測する状態予測部とを備えた第1態様の予測型行動決定装置を提供する。

【0027】

本発明の第12態様によれば、予測型行動決定装置において、自己の行動を決定する方法として、所定の環境について状態を観察して、状態値を取得し、取得した状態値に基づいて、前記環境の将来の状態変化を予測し、予測結果を基にして、行動決定のために最も適した将来の状態を目標状態として決定し、決定した目標状態を基にして、自己の行動を決定する行動決定方法を提供する。

【0028】

本発明の第13態様によれば、予測する状態変化は、前記環境の将来の自己の行動に影響されない状態変化である第12態様の行動決定方法を提供する。

【0029】

本発明の第14態様によれば、目標状態の決定を、予測した各将来状態に係る状態価値を参照して行う第12態様の行動決定方法を提供する。

【0030】

まず、本発明に関する基本的な概念について説明する。

【0031】

図1は課題の例を示す図である。図1では、座標 $(0, 0) - (1, 1)$ の空間1において、ボールBが直進運動をしている。空間1の上下左右の壁に当たったとき、ボールBは反射する。パドルPは左右方向にのみ動くことができるものとする。ボールBの状態は、位置座標 (B_x, B_y) と進む方向 B_t によって表現され、パドルPの状態は、位置座標 (P_x, P_y) によって表現される。ただし、 $P_y = 0$ で固定されている。

【0032】

各時刻ステップにおいて、パドルPの操作MPとして、 $\{LEFT$ (左に動く)、 $RIGHT$ (右に動く)、 $STAY$ (そのまま) $\}$ のうちいずれか1つを選択する。そして、ボールBをパドルPによって受けることができたとき、正の報酬が得られるものとし、一方、ボールBをパドルPによって受けることができなかったとき、負の報酬が得られるものとする。そしてここでのタスクは、得られる報酬をより多くすることである。

【0033】

これを行動決定問題として見た場合、各ステップごとに、入力としてボールBおよびパドルPの状態 (B_x, B_y, B_t, P_x) が与えられたとき、より多くの報酬が得られるように、パドルPの操作MPを選択する、ということになる。

【0034】

このような課題に対して、例えば、状態価値および行動基準を用いて、行動を決定する方法がすでに知られている。このような方法を、本願明細書では「政策 (Policy) に基づく行動決定」と呼ぶ。

【0035】

図2は政策に基づく行動決定に用いられる状態価値と行動基準の例を示す図である。ここでの「状態価値」は、外部の状態を評価して得た価値のことである。図2の例では、図1の空間1が 8×8 のセルに分割されており、各セルには、ボールBがその位置にあるときの状態価値が示されている。例えばボールBがセルCLの位置にあるとき、状態価値は「3」である。図2の例では、ボールBが空間1内の下面に到達したときに報酬がもらえるか否かが決定されることから、下面に近い位置ほど状態価値が高くなっている。状態価値は、事前に与えられたり

、学習によって獲得されたりする。

【0036】

また、各セルには、その位置に来たときどの行動を取るべきか、という行動基準が記述されている。図2の例では、図1と対応させて考えると、パドルPがボールBの下あたりにあり、ボールBが左下に向かって落ちてきているときに、ボールBの位置に対応したセルCLにおける行動基準として、左に動く行動に対して0.5、動かない行動に対して0.4、右に動く行動に対して0.1、という値が割り当てられている。この値を基にして、例えば最大値をもつ行動を選択したり、行動選択確率として計算したりして、最終的な行動が決定される。例えば最大値をとる行動を選択する場合には、セルCLでは、最大値0.5を持つ「左に動く」が選択される。

【0037】

このように、政策に基づく行動決定では、現在のボールBの状態（位置）に対応する状態価値が参照され、その状態に対する行動基準によって行動が決定される。すなわち、行動が現在の状態を考慮して決定されるので、例えばボールBの速度が早い場合など、事前に行動を決定する必要があるような場合等には対応できなかった。

【0038】

これに対して、本発明では、「予測に基づく行動決定」を行う。すなわち図3(a)に示すように、現在のボール位置B1に対して、複数のステップにわたる将来のボール位置を予測し、最も行動決定に適した目標状態となるボール位置B2を決定し、この目標状態を基にして行動決定を行う。

【0039】

図3(b)は予測された各ボール位置に対する状態価値が格納されたテーブル、図3(c)は予測ステップ数に対する状態価値の変化を示すグラフである。図3(c)から分かるように、予測ステップ数を増やすにつれて、ボールBは下面に向かって移動することになり、これとともに状態価値は徐々に増加する。そして、下面に到達してからはボールBは上方に向かうため、状態価値は減少に転じる。このため、状態価値は、予測ステップ数5のとき、最大値「8」になる。こ

ここで、例えば最大の状態価値を取った予測ステップ数と予測状態を行動決定に用いるものとする、位置B2が目標状態となる。これは、ボールBの状態でいうと、一番下面に近いときが目標状態として設定されることに相当する。このように、将来の、自己の行動に影響されない状態変化を予測して、予測した状態の中で行動決定のために最も適したものを目標状態として設定して行動決定を行うことによって、現在の行動をよりの確に決定することができる。

【0040】

以下、本発明の実施の形態について、図面を参照して説明する。なお、以下の説明では、図1の課題をタスクとして扱うものとする。図1の課題は、ピンポンのようにボールが移動してそれを跳ね返すと報酬が得られるものであるが、これは、サッカーゲームにおいてボールを取りに行く行動、移動ロボットが接近物を回避する行動等と同様の設定とみなすことができる。接近物の回避では、接近物の動きを予測してこれを回避することを目標状態として設定することによって、行動を決定できる。この他にも、ロボットが物を受け取る際の行動決定なども、同様の課題となる。

【0041】

(第1の実施形態)

図4は本発明の第1の実施形態に係る予測型行動決定装置10の構成を示す。この予測型行動決定装置10は、所定の環境11について状態を観察し、環境11の、自己の行動に影響されない状態変化を予測した上で、目標状態を定め、環境11に対する行動を決定する。

【0042】

状態観察部12は環境11の状態を観察し、現在の状態を表す状態値を取得する。ここでは、状態値を $s(t)$ と表す。図1の課題については、ボールBおよびパドルPの状態(B_x , B_y , B_t , P_x)が状態値 $s(t)$ として得られる。

【0043】

環境予測部13は状態観察部12によって取得された状態値 $s(t)$ に基づいて、環境11の将来の状態変化を予測する。ここでは、ボールBの座標(B_x ,

B y) を状態 s' として予測するものとする。すなわち、ボール B は基本的には直進し、壁で反射するので、将来の座標は、現在のボール B の座標と角度から解析的な計算によって予測することができる。またこの予測には、パドル P の操作 MP を考慮する必要はない。すなわち、当該装置 10 の行動に影響されない、将来の状態変化を予測することになる。本実施形態では、環境予測部 13 はこのような予測機能を予め備えているものとする。なお、このような予測機能は学習等によって得ることも可能であり、これについては後述する。

【0044】

状態価値記憶部 14 は図 3 (b) に示すように、各状態 s' すなわちボールの位置 (B_x , B_y) について状態価値をそれぞれ格納している。ここでは、状態 s' に対する状態価値を $V(s')$ と記述する。本課題では、ボール B が下の壁に当たるとき、報酬がもらえるか否かが判定されるので、下の壁に近い位置ほど、より高い状態価値が設定されている。なお、状態価値は全ての状態について設定されているのが好ましいが、これは実際には困難なので、例えば特定の部分状態にのみ状態価値を設定しておいてもよい。

【0045】

また、状態 s' の要素にボール B の進む方向 B_t を含めてもよく、この場合は、状態価値記憶部 14 に、(B_x , B_y , B_t) について状態価値をそれぞれ格納しておけばよい。

【0046】

目標状態決定部 15 は、環境予測部 13 による予測結果を基にして、行動の決定のために最も適した将来の状態を目標状態として決定する。ここでは、環境予測部 13 によって複数ステップにわたって予測された将来の状態 $s'(t + \Delta)$ について、それぞれ、状態価値記憶部 14 を参照して状態価値 $V(s'(t + \Delta))$ を求め、求めた状態価値から目標状態を決定する。

【0047】

第 1 の行動決定部としての予測に基づく行動決定部 16 は、目標状態決定部 15 によって決定された目標状態に対して、現在どのような行動を取るべきかを決定する。ここでは、ボール B の座標とパドル P の座標が一致したとき報酬がもら

えるので、目標状態から、パドルPがどの位置に向かえば報酬をもらえる可能性が高くなるを識別し、その位置に向かうように、操作MPを決定する。

【0048】

以下、図5のフローチャートを参照して、図4の予測型行動決定装置10の動作を説明する。まず、状態観察部12によって、環境11から、現在の状態を表す状態値 $s(t)$ が取得される(S11)。そして、何ステップ先の予測を行うかを変数 Δ で定義し、その初期値として1を与える(S12, S13)。なお、ステップS17で所定の条件を満たすまで、または、 Δ が所定値 n を超えるまで(S19)、 Δ をインクリメントしながら、以下のステップS14～S16を繰り返し実行する。

【0049】

ステップS14において、環境予測部13は、現在状態から Δ ステップ先の状態 $s'(t+\Delta)$ を予測する。例えば、ボールBが壁に当たらないときは、 B_x , B_y が1ステップ分変化し、方向 B_t は変化しない。そしてステップS15において、状態価値記憶部14から、ステップS14で予測された状態 $s'(t+\Delta)$ の状態価値 $V(s'(t+\Delta))$ が抽出される。

【0050】

ステップS16において、目標状態決定部15は、状態価値記憶部14から出力された状態価値 $V(s'(t+\Delta))$ を評価する。ここでは、評価条件として、所定の閾値を超えているか否かを判定する。そして、条件を満たす、すなわち状態価値が所定の閾値を越えていると判定したときは(S17でYES)、ステップS18に進み、状態 $s'(t+\Delta)$ を目標状態として、予測に基づく行動決定部16に与える。

【0051】

一方、条件を満たさないときは(S17でNO)ステップS19に進み、予測ステップ数の判定を行う。すなわち、 Δ が所定値 n を下回るときは(S19でNO)ステップS13に戻り、 Δ をインクリメントした後、同様の処理を行う。一方、 Δ が所定値以上のときは(S19でYES)ステップS1Aに進み、目標状態が決定できなかった旨を予測に基づく行動決定部16に通知する。

【0052】

予測に基づく行動決定部16は、目標状態決定部15からの出力を受けて、次の行動を決定する。しかし、目標状態が決定できなかった通知を受けた場合は、例えばランダムに、行動を決定する。

【0053】

このように本実施形態によると、環境11の状態変化の予測結果から、状態価値を参照して、行動決定のために最も適した将来の状態が目標状態として決定される。そして、この目標状態を基にして、自己の行動が決定される。このため、従来の将来を予測した行動決定よりも、行動決定の精度が格段に向上する。また、従来技術のように状態と行動との関係を予め記述していなくても、行動が決定できるので、簡易な構成により、様々な未経験の事態に対しても行動決定可能になる。また、環境11の将来の、当該装置10の行動に影響されない状態変化を予測しているので、1ステップ先だけではなく長い将来にわたっての予測も比較的容易に、精度良く、行うことができる。

【0054】

なお、本実施形態では、状態価値が所定の閾値を越えたときの状態を目標状態として決定するものとしたが、目標状態の決定は、これ以外にも、様々な方法が考えられる。例えば、予測したステップ内で状態価値が最大となった状態や、状態価値が極大になった状態を、目標状態として決定してもよい。また、前ステップとの状態価値の差分が所定値よりも小さくなったときの状態を目標状態として決定してもよい。あるいは、現在からのステップ数に応じて、状態価値を割り引いて評価するようにしてもよい。

【0055】

なお、本実施形態では、状態価値がテーブル形式で記憶されているものとしたが、この代わりに、ニューラル・ネットワーク等の関数近似手法を適用してもよい。この場合、現在の状態が入力されたとき、報酬の期待値が出力されるように学習がなされている必要がある。

【0056】

(第2の実施形態)

図6は本発明の第2の実施形態に係る予測型行動決定装置10Aの構成を示すブロック図である。図6において、図4と共通の構成要素には図4と同一の符号を付している。

【0057】

まず、価値変換部21について説明する。価値変換部21は、状態価値記憶部14Aが当該装置10Aの行動によって変化する状態を含む状態に係る状態価値を記憶しているとき、環境予測部13によって予測された、当該装置10Aの行動によって変化する状態を含まない将来状態 $s'(t+\Delta)$ の状態価値 $V(s'(t+\Delta))$ を、状態価値記憶部14Aに格納された状態価値を基にして求め、目標状態決定部15Aに与える。本実施形態では、状態価値記憶部14Aは、パドルPの位置 P_x も含めた状態 (B_x, B_y, B_t, P_x) についてそれぞれ状態価値を記憶しているものとし、環境予測部13からは将来状態 (B_x, B_y, B_t) が出力されるものとする。

【0058】

図7のフローチャートを参照して、価値変換部21の動作を説明する。まず、環境予測部13によって将来状態 $s'(t+\Delta)$ が予測されたとき、状態価値記憶部14Aをこの状態 $s'(t+\Delta)$ によって検索する(S21)。そして、3個の要素 (B_x, B_y, B_t) が一致する状態について、その状態価値の組 $V'(s'(t+\Delta))$ を抽出する(S22)。ここで、パドルPの座標 P_x が8通りであるとする、8個の状態価値が探索結果として出力される。

【0059】

そして、ステップS22において抽出された状態価値を比較し、その最大値を求め(S23)、得られた最大値を状態 $s'(t+\Delta)$ の状態価値 $V(s'(t+\Delta))$ として目標状態決定部15Aに出力する。

【0060】

また、第1の実施形態では、状態価値記憶部14に格納された状態価値は事前に与えられるものとしたが、本実施形態では、強化学習と呼ばれる手法によって自動的に学習させるものとする。「強化学習」とは、参考文献1(サットン(R. S. Sutton), バート(A. Barto)著, 「強化学習入門(Reinforcement Learnin

g: An Introduction) 」、(米国)、A Bradford Book, The MIT Press, 1998年3月)で知られる学習方法の1種であり、環境からの報酬信号と試行錯誤によって学習を行う手法であり、報酬の最大化を目的として行動決定を行う学習アルゴリズムのことである。

【0061】

本実施形態では、強化学習の中で、actor-critic学習という手法を利用する(参考文献1のp. 151~153参照)。actorは、どの状態でどの行動を取るべきかを記述する行動決定方策policyを持ち、状態sで、とりうる行動a1~anに対して選択確率が算出されて決定される。criticのVには状態価値と呼ばれる、ある状態でどれぐらい報酬がもらえそうかの期待値を示す値が格納される。状態価値は、報酬から計算されるTD誤差によって更新される。TD誤差 δ は、

$$\delta = r(t+1) + \gamma (V(s(t+1)) - V(s(t)))$$

で計算される。ここで $r(t+1)$ は報酬、 γ は割引率(discount rate)と呼ばれている。このTD誤差によって、critic内部の状態価値テーブルV(s)と、actorの持つPolicyの更新を行う。

【0062】

actor-critic法を利用して、状態価値記憶部14Aに記憶される状態価値を更新することによって、状態価値が不明な状態であっても、予測によって行動を決定することができる。状態価値更新部22はcriticテーブル更新則によって、状態価値記憶部14A内に記憶された状態価値を更新する。

【0063】

図8は強化学習によって学習された状態価値記憶部14Aの状態価値に基づき、価値変換部21が求めた状態価値の変化を示すグラフである。図8において、縦軸は算出された状態価値、横軸は予測のステップ数である。図8のグラフでは、予測ステップ数の増加に伴って状態価値が増加しており、80ステップあたりでピークを迎えている。一般に、状態価値は、図8のように、未学習領域から既学習領域に入ると増加し始め、その後に減少に転じる。この減少に転じるタイミングが、現在状態から予測されるもっとも意義のある状態、すなわち目標状態に

相当すると判断できる。

【0064】

目標状態設定部15Aは、第1の実施形態と同様に動作する。ただし、目標状態を決定できたか否かを示す選択信号を、行動選択部24に与える点が異なっている。予測に基づく行動決定部16Aは、第1の実施形態と同様に動作して行動を決定し、決定した行動を第1の行動候補として行動選択部24に出力する。

【0065】

第2の行動決定部としての政策に基づく行動決定部23は、所定の行動基準としての、強化学習におけるactor-critic手法によって学習された政策に基づいて行動を決定し、決定した行動を第2の行動候補として行動選択部24に出力する。

【0066】

行動選択部24は、予測に基づく行動決定部16Aから受けた第1の行動候補と、政策に基づく行動決定部23から受けた第2の行動候補とのうち、いずれか一方を、環境11への実際の行動として選択する。この選択には、目標状態決定部15Aから受けた選択信号を利用する。すなわち、選択信号が目標状態を決定できなかったことを示すときは、予測に基づく行動決定部16Aから受けた第1の行動候補は意味がないと考えて、政策に基づく行動決定部23から受けた第2の行動候補を実際の行動として選択する。そうでない場合は、予測に基づく行動決定部16Aから受けた第1の行動候補を実際の行動として選択する。

【0067】

このように本実施形態によると、状態価値を強化学習によって求めることによって、あらかじめ与えることが困難であった状態価値が自律的に獲得され、予測型行動決定が容易に実現できる。また、学習装置としての観点から本装置を見た場合は、報酬が与えられる時点を予測することになり、初期の実学習の状態が多い場合にも行動の指針を得ることができ、学習の効率が向上する。

【0068】

以下、従来の強化学習 (Reinforcement Learning; 以下、「RL」と記載) と、本実施形態に係る学習方法 (Prediction-based Reinforcement Learning; 以

下「PRL」と記載)のシミュレーションによる比較結果をいくつか示す。このシミュレーションでは、ボールが打てたときの報酬は1.0、打てなかったときの報酬は-1.0、またパドルが右または左に動いたときの報酬は-0.01とした。また、PRLでは、最初に状態価値記憶部14A内の状態価値をある程度作成するために3000試行(ボールが下面に当たるまでを1試行といい、以下epochと記載)は、RLと同様の処理で行動決定を行った。各epochにおけるボールの初期位置はランダムに設定した。

【0069】

図13はRLとPRLそれぞれの累積報酬の変化を表すグラフである。横軸はepoch数、縦軸は累積報酬である。最初の3000epochでは、両手法の累積報酬にはほとんど差はない。これは、この期間では、PRLはRLと同じ行動決定方法と学習法で動いているからである。学習の3000epoch以降では、PRLの方が、環境予測部13と予測に基づく行動決定部16Aとの利用によって、RLよりも良い性能を示している。これは、PRLでは、まだ学習が収束していない(すなわち、状態価値記憶部14A内の状態価値が完全には生成されていない)ときでも、予測によって適切な行動が決定できるので、ボールを打ち返す確率がRLよりも高くなるからである。また、PRLはボールを打ち返す位置の予測に基づき行動を決定しているため、不必要な行動をする割合が低くなっていることもその原因の一つと考えられる。

【0070】

図14は学習100epoch毎に評価を行った結果を示すグラフである。横軸はepoch数、縦軸は100epochごとに行われた性能評価の結果であり、ボールが打てた割合である。ここでのRLの行動決定規則には、ソフトマックス(softmax)と呼ばれる行動戦略を用いている。これは、各状態の各行動に付与された値に応じて確率的に行動決定を行う方法である。初期の3000epochの学習後、PRLmodelのパフォーマンスが、RLと比べて大きく向上している。この結果は、まだボールを打ち返す付近の状態しか学習が進んでいない(すなわち、状態価値が決定されていない)RL学習の初期段階に、PRLが環境予測部13と予測に基づく行動決定部16Aとの利用によって、学習が進行していない状態に対して

も適切な行動を決定できたことが大きな要因である。

【0071】

図15はRLの行動決定規則をgreedy戦略に変更した場合のグラフである。図14と比べて、RLの性能は全体的に見ると向上している。しかし、学習していない状態に評価時に遭遇すると性能は急激に落ちる。RLは、PRLに比べると、精度が悪く安定した動作を示していないことが分かる。

【0072】

なお、本実施形態では、状態価値をテーブルとして表現して強化学習により学習するものとしたが、ニューラル・ネットワークによって学習させることもできる。この場合、ニューラル・ネットワークの汎化能力によって未経験の状態に対する価値を出力することが期待される。しかし、この手法は、状態価値に不連続な状態が少ないときに有効であると考えられる。

【0073】

(第3の実施形態)

上述の実施形態では、予測に基づく行動決定部16、16Aの機能は予め与えられているものとしたが、行動生成機能を予め与えることが難しい場合は、目標状態に到達するための行動生成能力の獲得が必要になる。本実施形態では、このような行動生成能力を学習によって獲得させるものとする。

【0074】

図9は図6の構成における予測に基づく行動決定部16Aの本実施形態に係る内部構成を示す図である。図9において、31は状態値 $s(t)$ を受け、この状態値 $s(t)$ が表す現在状態から、その前ステップにおける状態と行動を検出する行動付状態変化検出部、32は行動付状態変化検出部31によって検出された現在状態 $s(t)$ 並びにその前ステップにおける状態 $s(t-1)$ および行動 $a(t-1)$ の組合せを状態変化として記憶する行動付状態変化記憶部、33は行動付状態変化記憶部32から、現在状態から目標状態までの期間の状態変化の履歴を検索し、この検索結果を基にして行動を決定する行動計画部である。

【0075】

図10のフローチャートを参照して、図9の予測に基づく行動決定部16Aの

動作を説明する。図10(a)は状態変化を蓄積する場合の動作、図10(b)は行動計画を行う場合の動作を示す。この2つの動作は、同時並行して実行できる。

【0076】

状態変化を蓄積する場合、まず、環境11から状態 $s(t)$ を受ける(S31)。この現在の状態 $s(t)$ とそのときの行動 $a(t)$ はワーキングメモリに蓄えておく。そして、前ステップにおける状態 $s(t-1)$ と行動 $a(t-1)$ をワーキングメモリから取り出し(S32)、状態 $s(t)$ とともに行動付状態変化記憶部32に格納する(S33)。これは、状態 $s(t-1)$ のときに行動 $a(t-1)$ をとったら状態 $s(t)$ に変化したという、行動に起因する状態変化を表している。

【0077】

また、行動計画を行う場合、まず、目標状態決定部15Aから送られてきた目標状態を探索したい状態 $x_s(n)$ として設定する(S34)。そして、探索したい状態 $x_s(n)$ を行動付状態変化記憶部32から探索し(S35)、検出されたときは(S36でYES)、探索したい状態 $x_s(n)$ と対になって記憶されている状態 $x_s(n-1)$ および行動 $x_a(n-1)$ をワーキングメモリに格納する(S37)。その後、ステップS38に進み、探索したい状態の1ステップ前の状態 $x_s(n-1)$ が現在の状態 $s(t-1)$ でなければ(S38でNO)、探索したい状態を更新し(S39)、ステップS35に戻る。同様の処理を繰り返し実行し、ステップS38において、探索したい状態の1ステップ前の状態 $x_s(n-1)$ が現在の状態 $s(t-1)$ と一致したとき(YES)、それまでにワーキングメモリに格納されていた状態 x_s と行動 x_a の系列を行動計画として出力する。

【0078】

一方、ステップS36において、探索したい状態 $x_s(n)$ が行動付状態変化記憶部32から検出できないときは(NO)、行動不能と判断して(S3A)、処理を終了する。なおこの場合は、予測に基づく行動決定部16Aから、正しい行動が決定できないという信号が出力され、行動選択部24は、政策に基づく行

動決定部 23 から出力された第 2 の行動候補を実際の行動として選択する。

【0079】

このような動作によって、現在の行動のみならず、現在状態から目標状態に至るまでの行動計画が得られるので、行動計画が一旦完了した後は、その行動計画に従って行動候補を順に出力すればよい。これにより、処理量が格段に少なくなるので、特に長期にわたる予測誤差が少ない場合は、好ましい。もちろん、毎ステップごとに、目標状態までの行動計画を算出し直してもよい。この場合は、予測が完全でない場合であっても、行動が決定できる。

【0080】

なお、本実施形態では、目標状態から現在状態に向かって後ろ向き探索を行うものとしたが、現在の状態 $s(t)$ と行動 $a(t)$ から $s(t+1)$ を算出する前向き探索を用いても、同様に行動計画を作成することができる。

【0081】

(第 4 の実施形態)

本実施形態では、学習によって状態予測を実現するものとする。

【0082】

図 11 は図 4 および図 6 の構成における環境予測部 13 の本実施形態に係る内部構成を示す図である。図 11 において、41 は状態値 $s(t)$ を受け、この状態値 $s(t)$ が表す現在状態から、その前ステップにおける状態を検出する状態変化検出部、42 は状態変化検出部 41 によって検出された現在状態 $s(t)$ およびその前ステップにおける状態 $s(t-1)$ を状態変化として記憶する状態変化記憶部、43 は状態変化記憶部 42 から、現在状態の後の状態を予測する状態予測部である。

【0083】

図 12 のフローチャートを参照して、図 11 の環境予測部 13 の動作を説明する。図 12 (a) は状態変化を蓄積する場合の動作、図 12 (b) は状態予測を行う場合の動作を示す。この 2 つの動作は、同時並行して実行できる。

【0084】

状態変化を蓄積する場合、まず、環境 10 から状態 $s(t)$ を受ける (S41

)。この現在の状態 $s(t)$ はワーキングメモリに蓄えておく。そして、前ステップにおける状態 $s(t-1)$ の組合せをワーキングメモリから取り出し (S42)、状態 $s(t)$ とともに状態変化記憶部 42 に格納する。これは、状態 $s(t-1)$ の次に状態 $s(t)$ になったという状態変化を表している。

【0085】

また、状態予測を行う場合、まず、環境 11 から取得した現在の状態 $s(t)$ を探索したい状態 $y_s(n)$ として設定する (S44)。そして、探索したい状態 $y_s(n)$ を状態変化記憶部 42 から探索し (S45)、検出されたときは (S46でYES)、探索したい状態 $y_s(n)$ と対になって記憶されている 1 ステップ後の状態 $y_s(n+1)$ を状態変化記憶部 42 から取り出し、出力する (S47)。その後、ステップ S48 に進み、目標状態決定部 15, 15A から評価信号による再予測依頼を受けたときは (YES)、探索したい状態を更新し (S49)、ステップ S45 に戻る。

【0086】

一方、ステップ S46 において、探索したい状態 $y_s(n)$ が状態変化記憶部 42 から検出できないときは (NO)、予測先不明と判断して (S4A)、処理を終了する。なおこの場合は、環境予測部 13 から、正しい予測ができないという信号が出力され、行動選択部 24 は、政策に基づく行動決定部 23 から出力された第 2 の行動候補を実際の行動として選択する。

【0087】

このような方法によって、環境予測部 13 の機能を予め作成しなくても、学習によって取得することができる。

【0088】

なお、状態変化記憶部 42 の学習のために、ニューラル・ネットワーク等の関数近似手法によって次の状態を予測させることもできる。この場合、ニューラル・ネットワークが本来持つ汎化能力によって、経験したことがない状態 $s(t)$ に対しても、適切な 1 ステップ後の状態 $s(t+1)$ を出力できる可能性がある。

【0089】

【発明の効果】

以上のように本発明によると、環境の将来状態を予測し、予測の結果決定された、行動決定のために最も適した目標状態を基にして、行動が決定されるので、将来の状態変化がより適切に考慮され、行動決定の精度が向上する。

【図面の簡単な説明】**【図 1】**

本発明、および各実施形態を説明するための課題を示す図である。

【図 2】

政策に基づく行動決定に用いられる状態価値と行動基準の例である。

【図 3】

本発明に係る、予測に基づく行動決定を示す図である。

【図 4】

本発明の第 1 の実施形態に係る予測型行動決定装置の構成を示すブロック図である。

【図 5】

図 4 の予測型行動決定装置の動作を示すフローチャートである。

【図 6】

本発明の第 2 の実施形態に係る予測型行動決定装置の構成を示すブロック図である。

【図 7】

図 6 の構成における価値変換部の動作を示すフローチャートである。

【図 8】

予測ステップ数に伴う状態価値の変化の一例を示すグラフである。

【図 9】

本発明の第 3 の実施形態における予測に基づく行動決定部の内部構成を示す図である。

【図 10】

図 9 の構成の動作を示すフローチャートである。

【図 11】

本発明の第 4 の実施形態における環境予測部の内部構成を示す図である。

【図 1 2】

図 1 1 の構成の動作を示すフローチャートである。

【図 1 3】

本発明の実施形態に係る学習方法と従来の強化学習のシミュレーション結果を示すグラフである。

【図 1 4】

本発明の実施形態に係る学習方法と従来の強化学習のシミュレーション結果を示すグラフである。

【図 1 5】

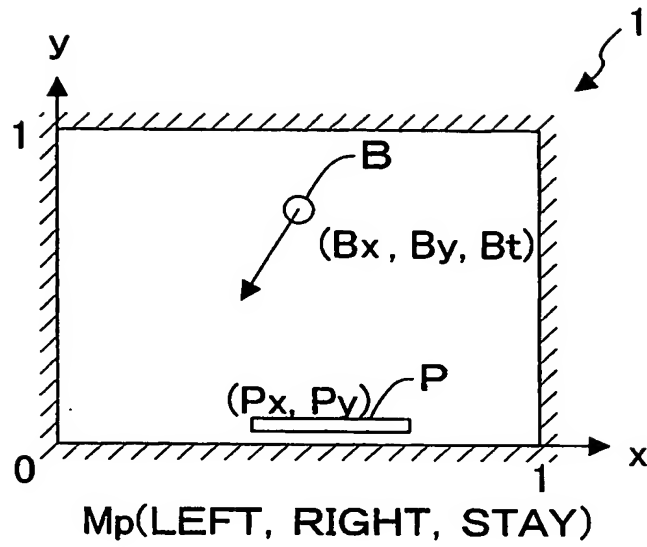
本発明の実施形態に係る学習方法と従来の強化学習のシミュレーション結果を示すグラフである。

【符号の説明】

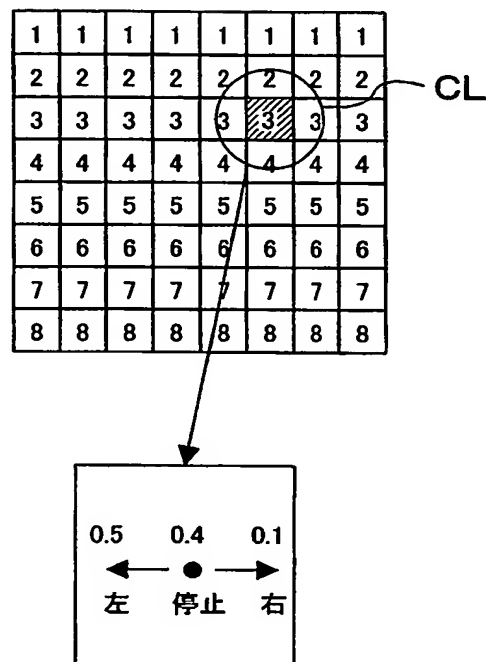
- 1 0, 1 0 A 予測型行動決定装置
- 1 1 環境
- 1 2 状態観察部
- 1 3 環境予測部
- 1 4, 1 4 A 状態価値記憶部
- 1 5, 1 5 A 目標状態決定部
- 1 6, 1 6 A 予測に基づく行動決定部（第 1 の行動決定部）
- 2 1 価値変換部
- 2 3 政策に基づく行動決定部（第 2 の行動決定部）
- 2 4 行動選択部
- 3 1 行動付状態変化検出部
- 3 2 行動付状態変化記憶部
- 3 3 行動計画部
- 4 1 状態変化検出部
- 4 2 状態変化記憶部
- 4 3 状態予測部

【書類名】 図面

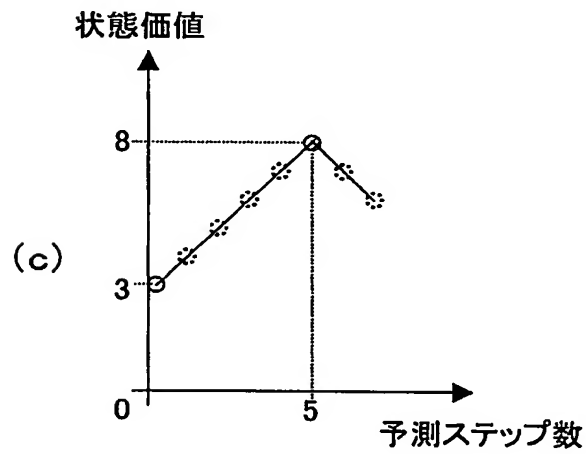
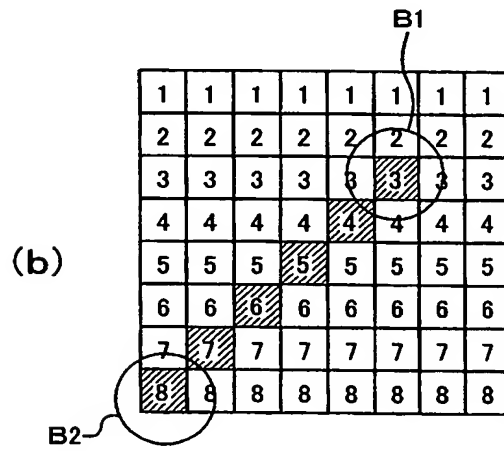
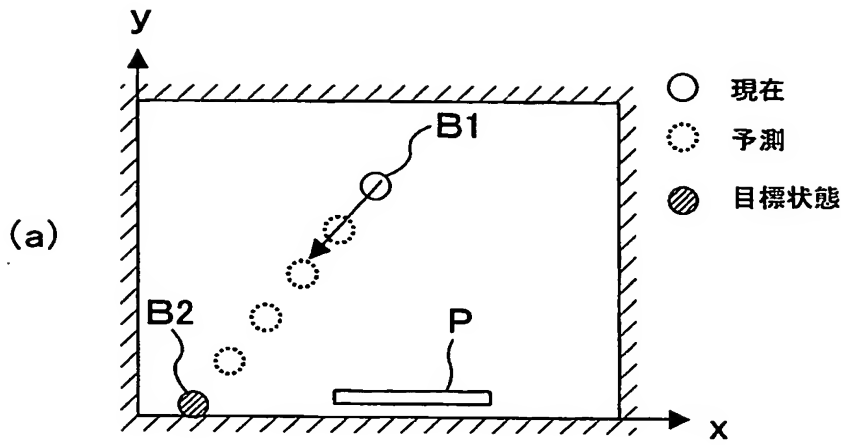
【図 1】



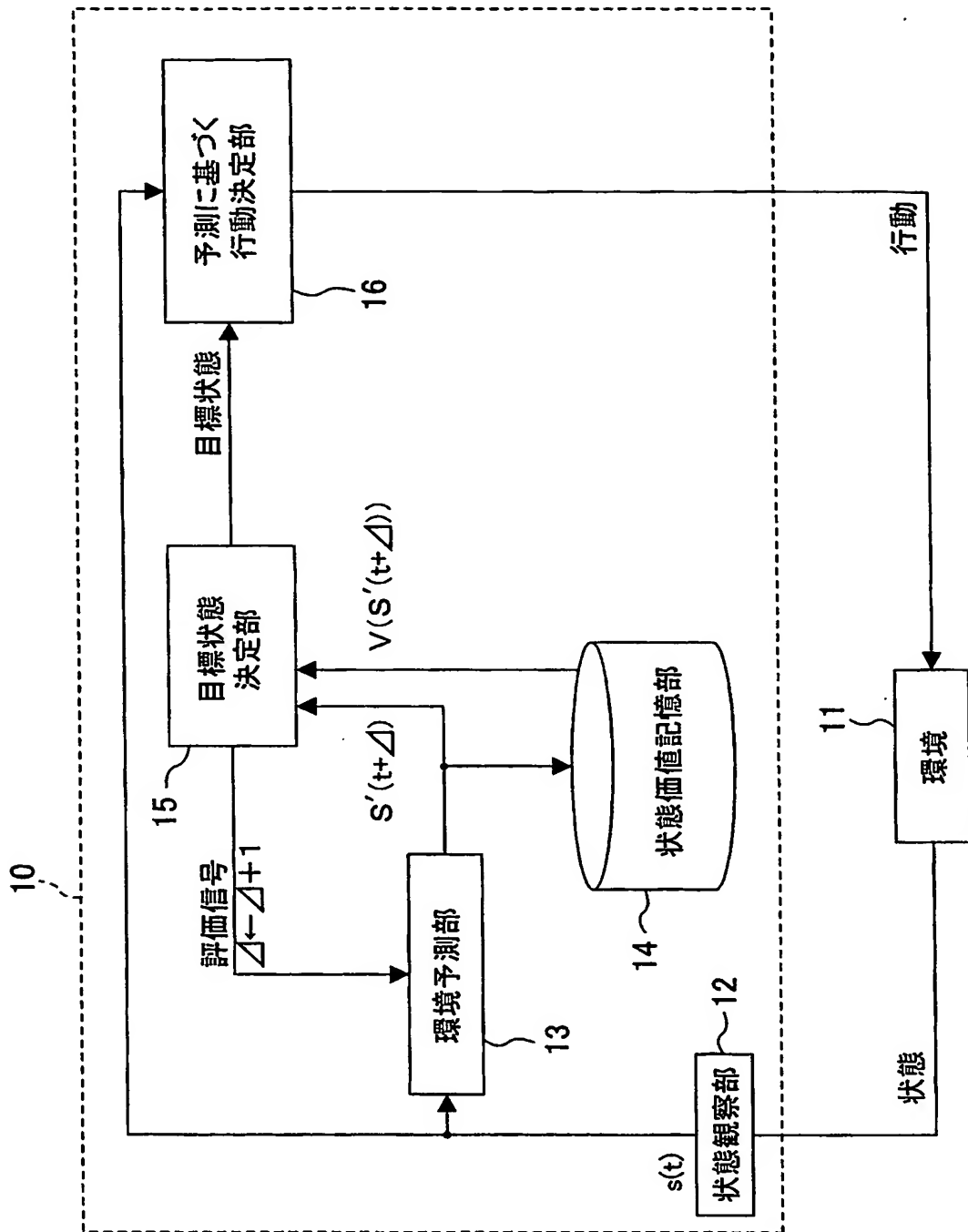
【図 2】



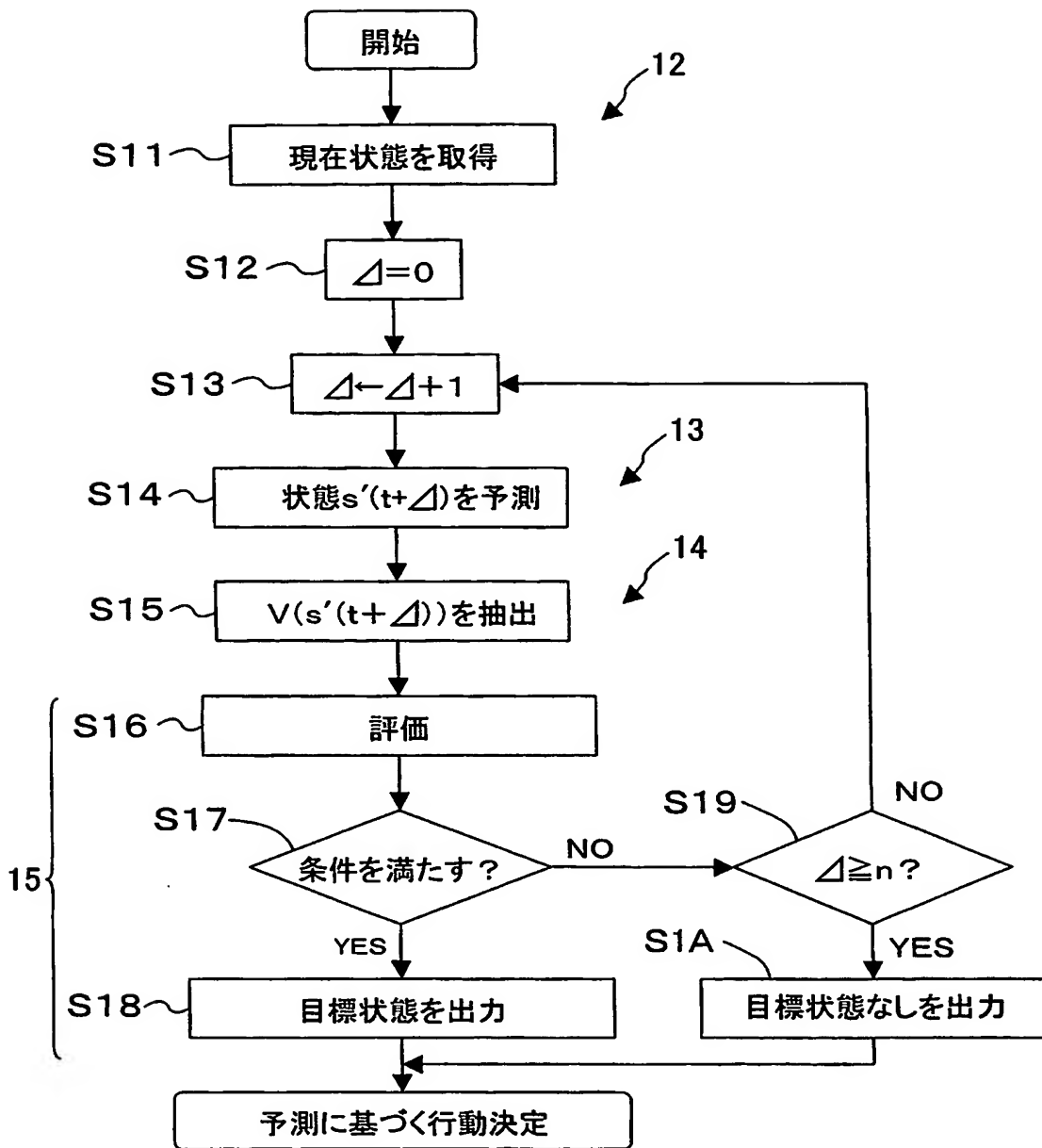
【図 3】



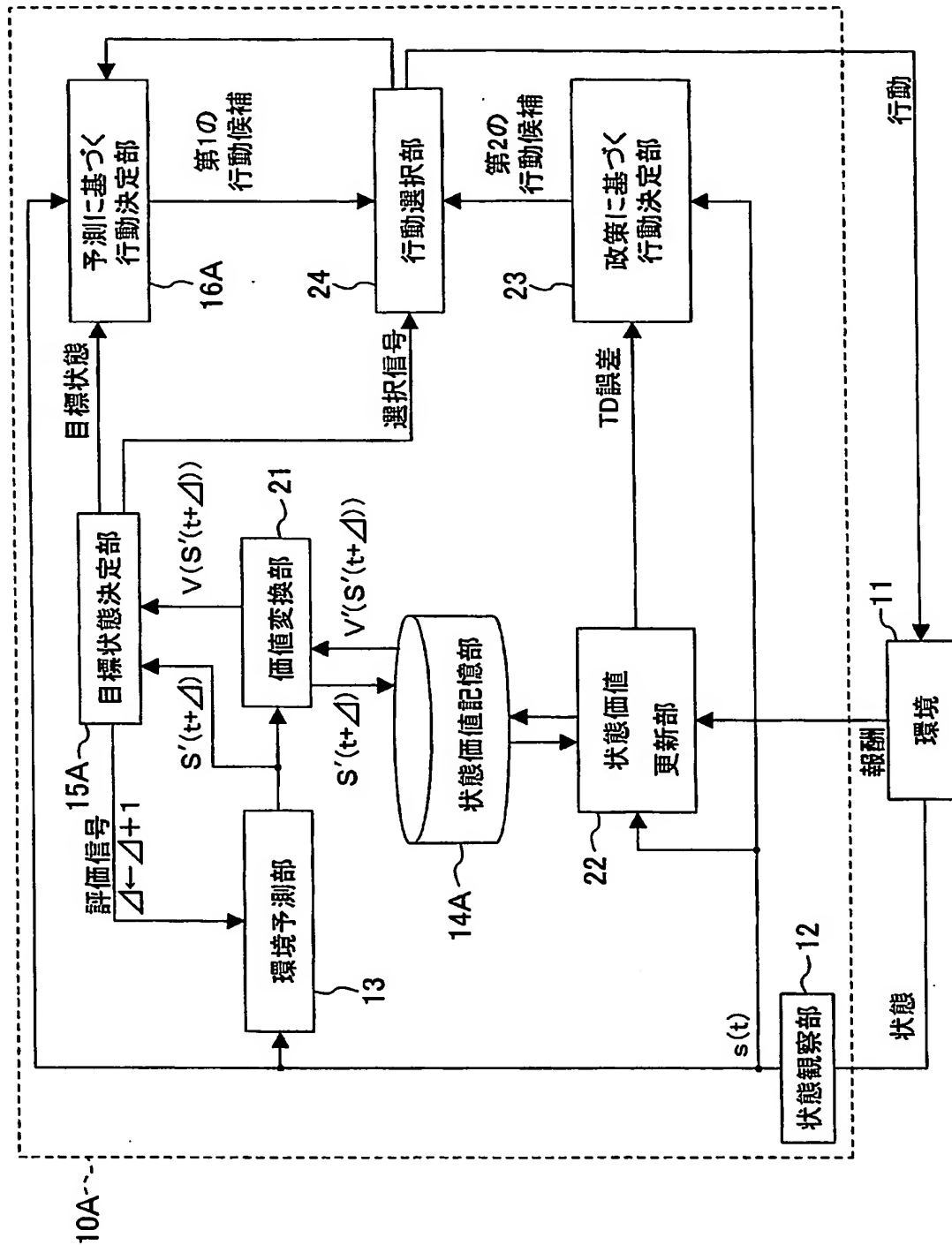
【図 4】



【図 5】

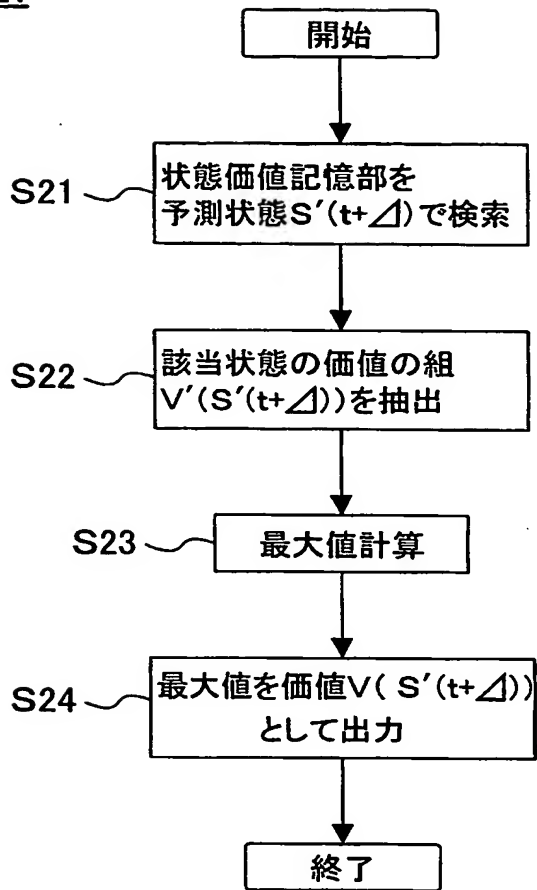


【図6】

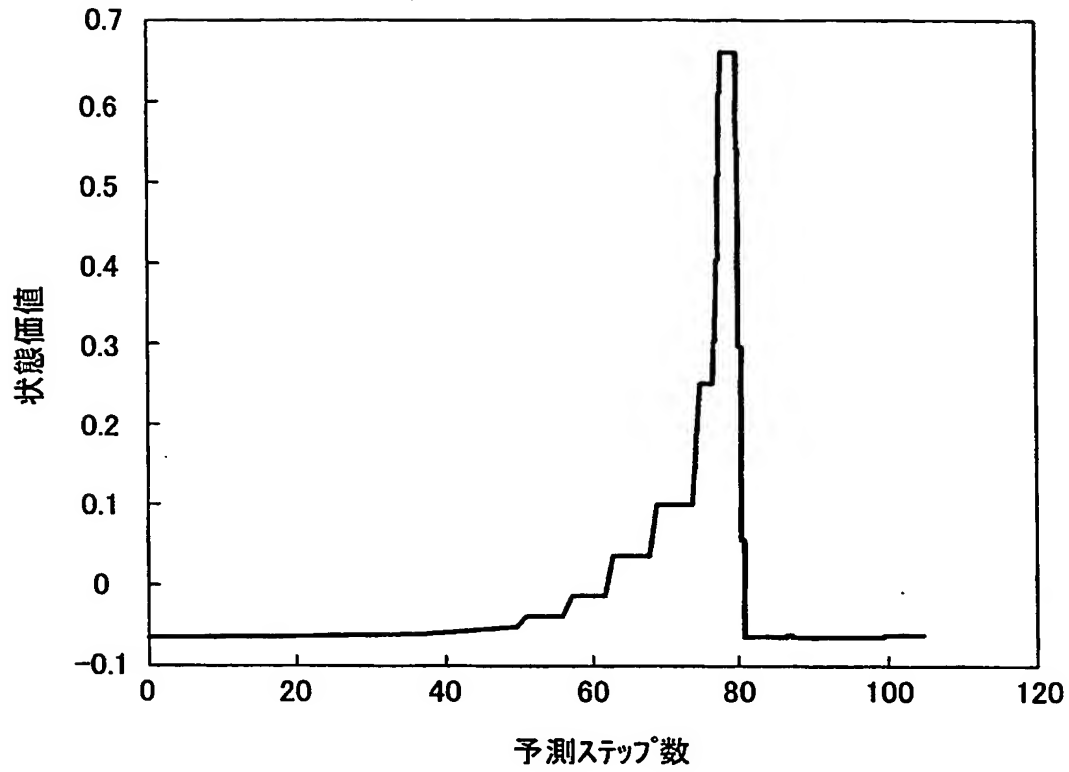


【図 7】

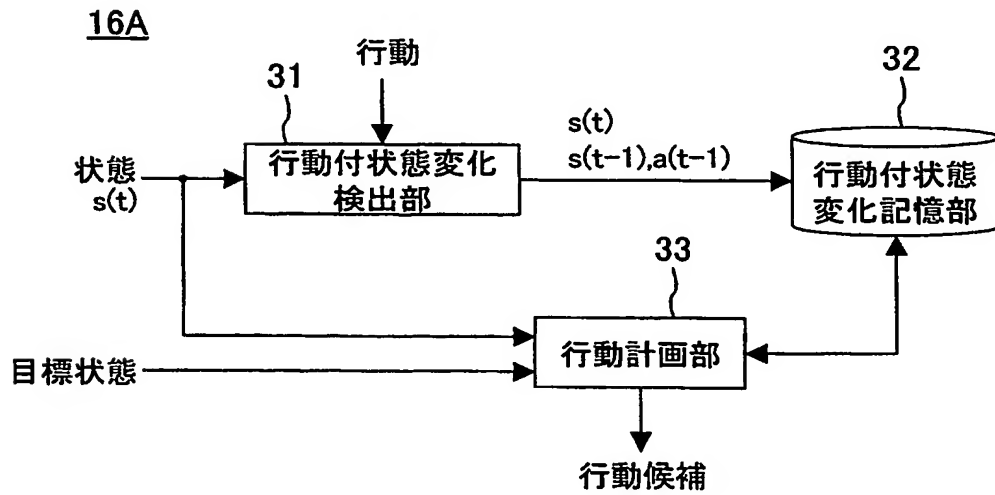
21



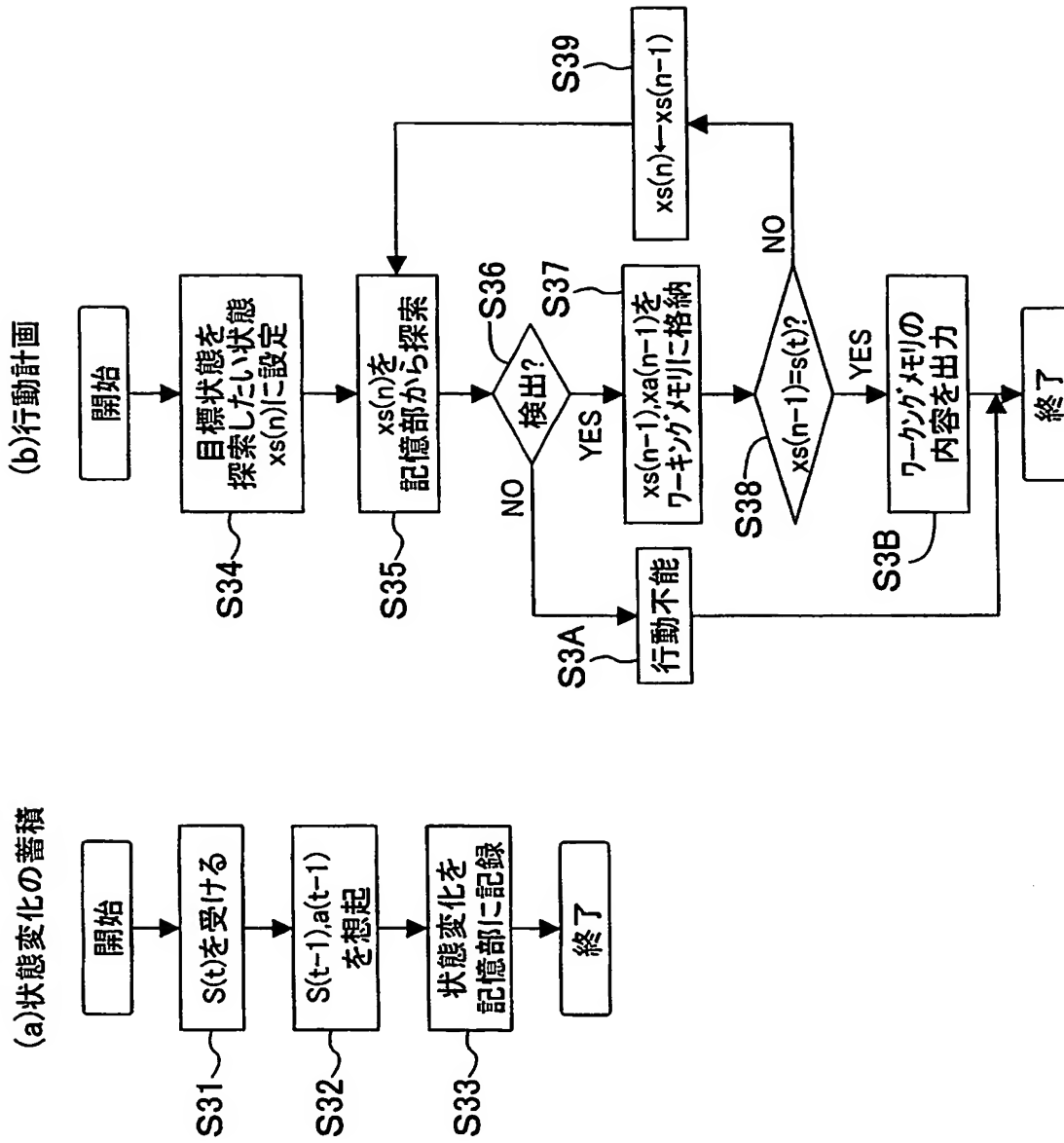
【図 8】



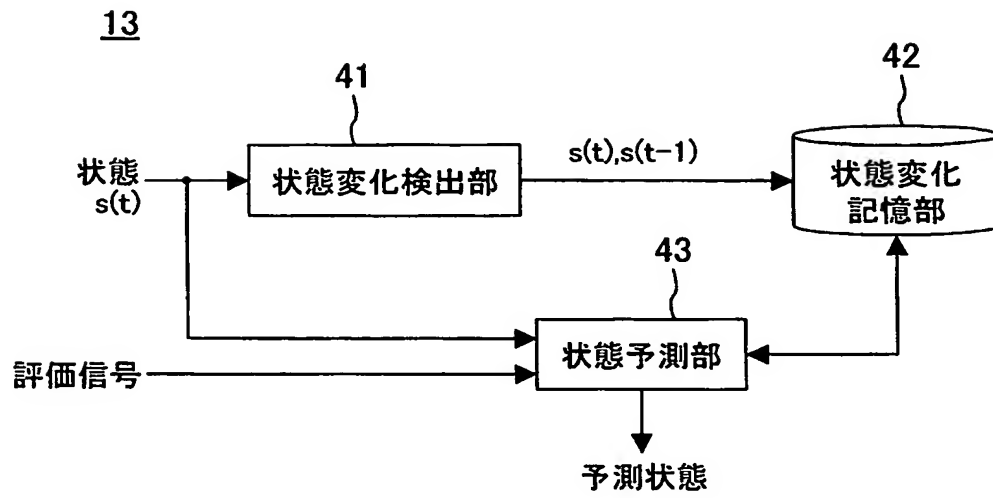
【図 9】



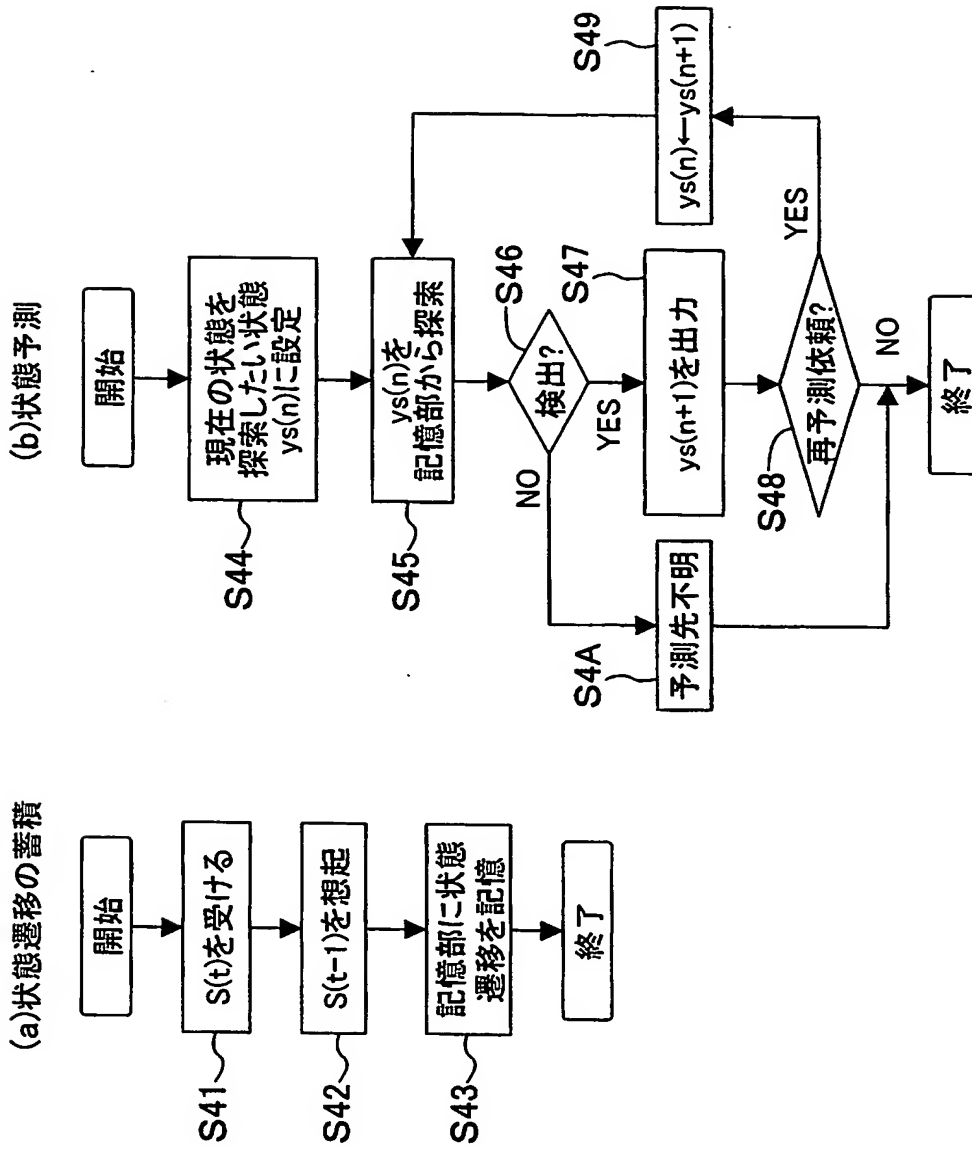
【図 10】



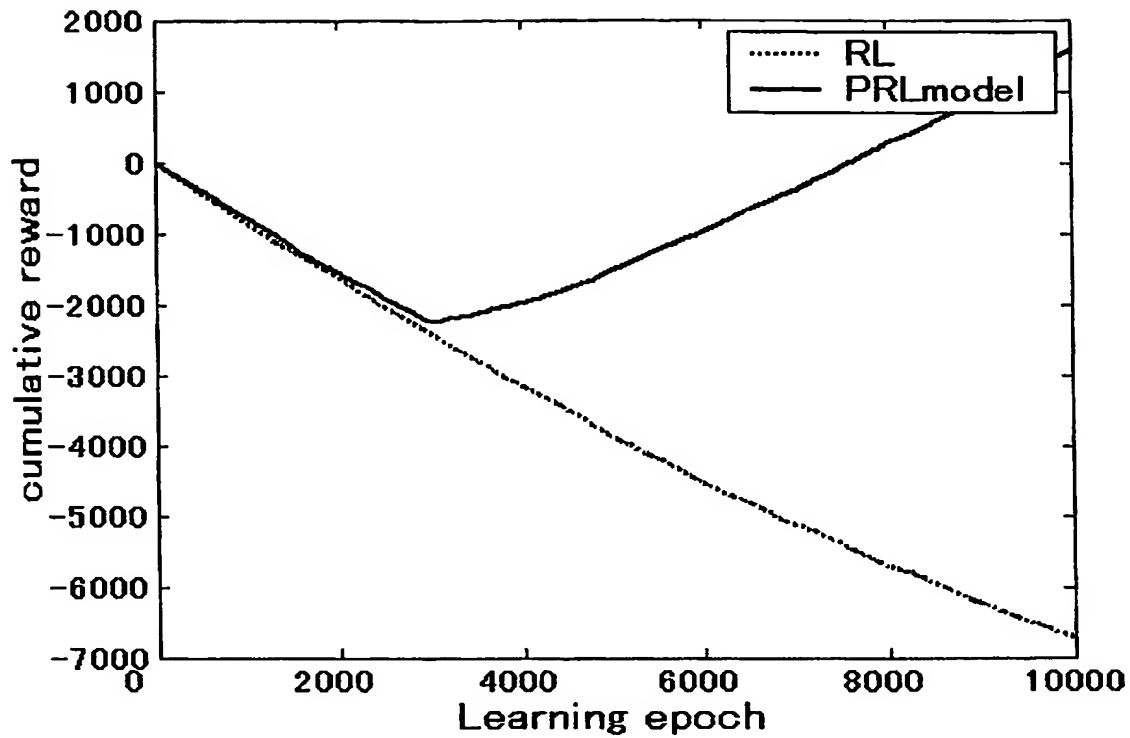
【図 11】



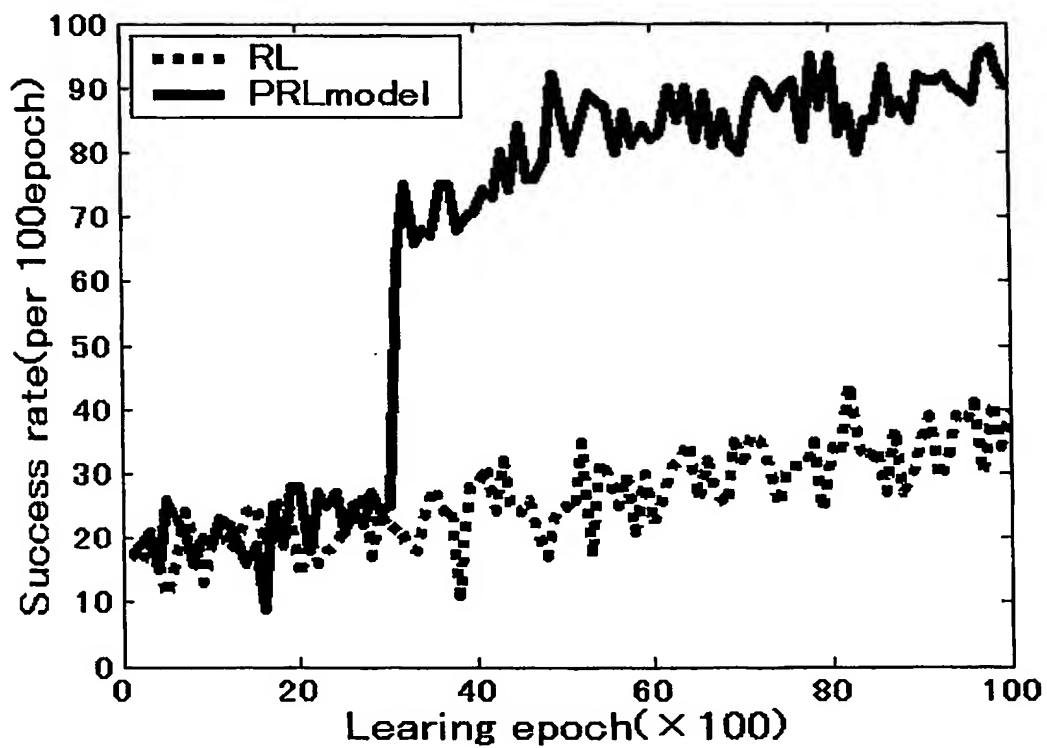
【図 12】



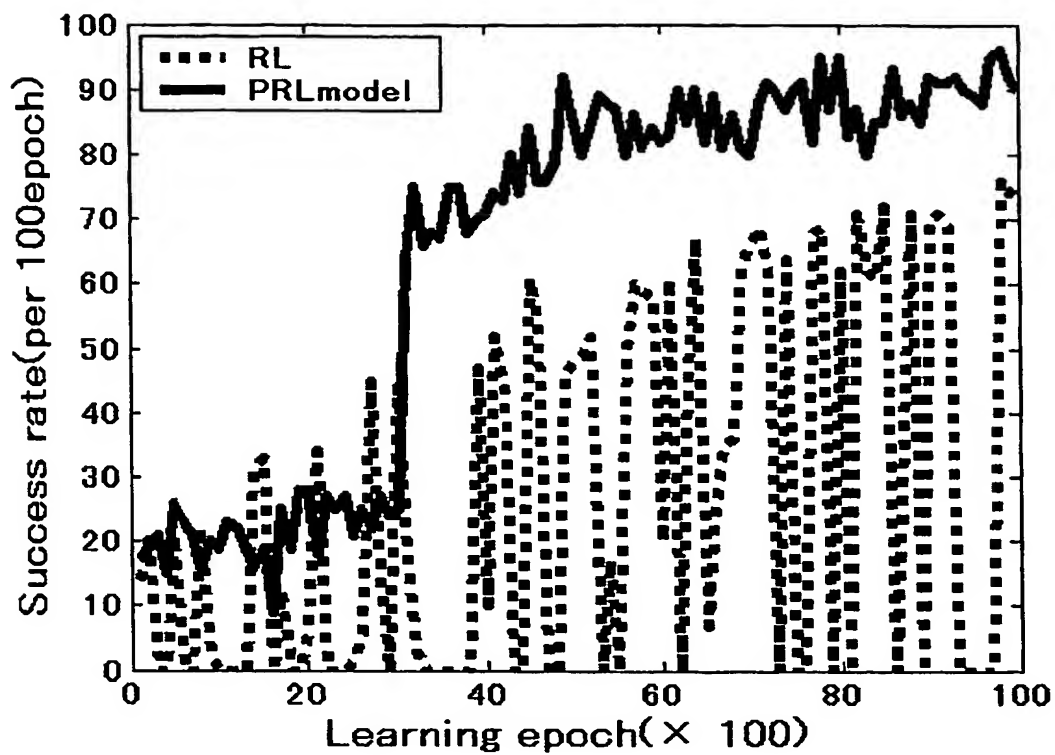
【図 1 3】



【図 1 4】



【図15】



【書類名】 要約書

【要約】

【課題】 予測型行動決定装置において、行動決定のために将来の状態予測をより適切に行い、行動決定の精度や能力を向上させる。

【解決手段】 予測型行動決定装置 10 において、状態観察部 12 は所定の環境 11 について状態を観察し、状態値 $s(t)$ を取得する。環境予測部 13 は取得された状態値 $s(t)$ に基づいて、環境 11 の将来の、装置 10 の行動に影響されない状態変化を予測する。目標状態決定部 15 は状態価値記憶部 14 を参照して、行動決定のために最も適した将来の状態を目標状態として決定する。予測に基づく行動決定部 16 は決定された目標状態を基にして、自己の行動を決定する。

【選択図】 図 4

特願 2 0 0 3 - 0 2 3 3 4 2

出 願 人 履 歴 情 報

識別番号

[0 0 0 0 0 5 8 2 1]

1. 変更年月日

1 9 9 0 年 8 月 2 8 日

[変更理由]

新規登録

住 所

大阪府門真市大字門真 1 0 0 6 番地

氏 名

松下電器産業株式会社